DOCUMENT RESUME

ED 130 495 FL 007 929

AUTHOR	Perry, Jessica, Ed.; Pietrzyk, Alfred, Ed.
TITLE	Preliminaries to the Design of LINCS Indexing Tools. LINCS Project Document Series.
INSTITUTION	Center for Applied Linguistics, Washington, D.C. Language Information Network and Clearinghouse System.
SPONS AGENCY	National Science Foundation, Washington, D.C.
REPORT NO	CALLINCS-69-15
PUB DATE	Jul 71
GRANT	NSP-GN-771
NOTE	55p.
EDRS PRICE	MF-$0.83 HC-$3.50 Plus Postage.
DESCRIPTORS	Automatic Indexing: Classification; *Indexing; *Information Networks; Information Processing; *Information Retrieval; Information Science; *Information Systems; Lexicography; *Linguistics; Search Strategies; Subject Index Terms; *Thesauri; *Vocabulary
IDENTIFIERS	*Language Sciences

ABSTRACT

The four chapters included in this report are based on LINCS project activities undertaken since 1968, with an emphasis on indexing tools in the language sciences and related problems. Chapter one, "Indexing Tools for the Language Sciences: Methodology," discusses the development of a LINCS thesaurus and its role in the LINCS network. Chapter two, "Vocabulary and Indexing for LINCS: Some Preliminary Considerations," discusses LINCS indexing procedures. Chapter three, "A Preliminary Classification for Language Sciences Information: Working Outline," discusses the requirements for a classification system which could constitute a framework for the LINCS thesaurus. Chapter four, "Vocabulary Control for the LINCS Reference Management System (RMS)," summarizes the initial indexing approaches and authority file management techniques which, at this time, are considered to be optimal for use in the proposed Reference Management System (RMS), the automated central clearinghouse and secondary processing facility of LINCS. (Author/AM)

CENTER FOR APPLIED LINGUISTICS

LANGUAGE INFORMATION NETWORK AND CLEARINGHOUSE SYSTEM (LINCS)

PRELIMINARIES TO THE DESIGN OF LINCS INDEXING TOOLS

Prepared and edited by

Jessica Perry

Alfred Pietrzyk

LINCS PROJECT DOCUMENT SERIES / NATIONAL SCIENCE FOUNDATION GRANT

CALLINCS-69-15 July 1971 NSF GN--771

CENTER FOR APPLIED LINGUISTICS, 1717 MASSACHUSETTS AVENUE, N.W., WASHINGTON, D.C. 20036

PRELIMINARIES TO THE DESIGN OF LINCS INDEXING TOOLS

Prepared and edited by

Jessica Perry

Alfred Pietrzyk

CONTENTS

NOTE

The four chapters included in this report are based on LINCS project activities undertaken since 1968 with an emphasis on indexing tools in the language sciences and related problems, some of which are also treated in the following documents of the project series:*

Lewis, Kathleen P., comp. Indexing tools and terminology sources in the language sciences: A bibliographical listing. LINCS #2-68, NSF GN-653. Washington, D.C.: Center for Applied Linguistics, 1968, 20 p. (ERIC: ED 021 245).

Pietrzyk, Alfred; Lamberts, Frances; Freeman, Robert R. File-management techniques and systems with applications to information retrieval: A selective bibliography. LINCS #3-68, NSF GN-653. Washington, D.C.: Center for Applied Linguistics, 1968, 27 p. (CFSTI: PB 178 792).

Rosenfeld, Samuel A.; Sable, Jerome. Requirements for LINCS file management system. LINCS #8-69, NSF GN-771. Washington, D.C.: Center for Applied Linguistics, 1969, paged by section. (CFSTI: PB 186 472).

Rappaport, Miriam. Citation patterns in selected core journals for linguistics. LINCS #13-69, NSF GN-771. Washington, D.C.: Center for Applied Linguistics, 1971, iii, 23 p.

Garvin, Paul L. Specialty trends in the language sciences. LINCS #16-69, NSF GN-771. Washington, D.C.: Center for Applied Linguistics, 1969, ii, 29 p. (ERIC: ED 034 983).

Ebersole, Joseph L. Some probable technological trends and their impact on an information network system. LINCS #3-70, NSF GN-771. Washington, D.C.: Center for Applied Linguistics, 1970, 13 p. (CFSTI: PB 192 494).

Zisa, Charles A. Language classification and indexing. With an annotated bibliography. LINCS #5-70, NSF GN-771. Washington, D.C.: Center for Applied Linguistics, 1970, 21 p.

Gifford, Carolyn. A survey of indexing tools for the language sciences. CALLINCS-70-6 , NSF GN-771. Washington, D.C.: Center for Applied Linguistics, 1971.

Rose, Priscilla. Linguistic bibliography count. LINCS #10-70 P, NSF GN-771. Washington, D.C.: Center for Applied Linguistics, 1971.

*Documents marked "P" are preliminary working papers for limited circulation only.

Chapter 1

INDEXING TOOLS FOR THE LANGUAGE SCIENCES: METHODOLOGY

By Jessica Perry

1. Introduction

With the support of the National Science Foundation, the Center for Applied Linguistics (CAL) has undertaken the responsibility of attempting to develop a viable information network to serve the users of language information.

Two questions immediately arise: 1) who are the users of language information, and 2) what is language information?

The first question cannot be answered definitively, of course, until there is some sort of information service for those who have serious information needs in the language sciences to use. One thing, however, seems to be clear: the users of the evolving Language Information Network and Clearinghouse System (LINCS) will not all be linguists. Many will be persons from outside the core discipline of linguistics who need linguistic information in connection with problems in other fields. At the same time LINCS will aim to serve the linguist effectively by giving him various information products that he now lacks or that are scattered among a wide variety of information sources of uneven quality and timeliness (see Part II of Freeman, Pietrzyk and Roberts [4]). These two uses of linguistic information are discussed in detail by Paul Garvin in Special Trends in the Language Sciences [5].

In this same report Garvin also discusses the question of the scope of linguistic information to be included in LINCS, and his chart on page 22 shows at a glance the relationships between "linguistics" and other fields as revealed in recent literature. The question of the scope of language information is obviously the "other side of the coin" of the question of the users, and the chart as well as the discussion by Garvin and others can serve as a guideline for the orderly growth and coverage of LINCS.

Using the guidelines given by Garvin and others as the conceptual framework of LINCS, we next have a series of problems connected with establishing a prototype LINCS in order to test the viability of the concept. These problems involve both operational and philosophical considerations such as the following: 1) How do we develop criteria for selection of input to LINCS, and how do we set up workable operational acquisition procedures for individual documents? 2) What will be the optimum indexing language to enable users with a wide variety of information needs to find relevant documents within the files of LINCS? 3) How will this index language interface with the various indexing languages of the ongoing information centers with which LINCS will be cooperating? 4) Especially considering the varying backgrounds of potential users of LINCS, where

will the responsibility for question analysis lie? Can the user translate his information need into the LINCS index terms, or will he be helped by the LINCS staff, or will they share this chore? Will the tolerable time of response of LINCS influence this decision? What effect will distance have upon it? Will the user be presented with the results of s small representative sample of responses on the basis of which he can revise his search, etc? To consider extremes, will the analysis take place in a traditional library situation, or in an on-line, real-time computer situation? 5) How will the files of LINCS be organized? Will the index terms for the documents of LINCS be stored to facilitate manual retrieval, or will they be stored for retrieval by some mechanized device or by computer? Will they be filed for sequential, document-index term searching, or will they be filed for so-called inverted, term-document searching? 6) What kinds of search strategy will be needed? Will the LINCS system, for example, need strategies other than those provided by the usual Boolean "and", "or", and "not", such as searches for quantities or ranges? How many index terms are likely to comprise the average search, the most complex search, i.e., number of index terms, the simplest search? How will the logic of the search formulation affect the cost and speed of searching? 7) Finally, how and in what form will the results of LINCS searches be disseminsted to users? Will LINCS provide on-demand retroapective searches, recurring current-awareness searches, or both? Will it attempt selective dissemination based on users' interest profiles? Will it issue specialized bibliographies or abstract bulletins based on topics of current interest? Will it send copies of original documents to its users? Will it attempt synthesis or stste-of-the-art- reports, such as various information center issues?

These are the major questions and problems facing any nascent information service. Each major problem offers a number of potentially feasible solutions. There are no solutions that are "right", across the board for all information services. Each service must work out its own solutions with respect to its own users, its own budget and the skill of its own staff. Information science has yet to come up with a way to predict solutions to these problems other than very careful, controlled testing of alternatives upon representative users. The Center for Applied Linguistics is prepared to carry out all necessary testing and evaluation of LINCS as it is developed and implemented, and to answer these questions as data on user needs are acquired.

2. The Indexing Language of LINCS

Aside from the production of the guidelines for establishing the scope and usership of language information, work which was absolutely fundamental as a basis upon which to build LINCS, recent efforts have been devoted to critical analysis of a number of available indexing languages (see Lewis [7] and Gifford [6]) and the choice of an index language for the first LINCS experiment, i.e.,

1) preparation of a small sample of "core" index terms,

2) indexing a small sample of "core" documents by means of these terms,

3) searches of the sample by a carefully selected group of representative users,

4) evaluation, refinement, and extension of the index language, the indexed file, and the user population followed by further searches and evaluation.

3. LINCS Thesaurus

In view of the following considerations, the thesaurus was chosen as the most advantageous index language for LINCS:

1) LINCS will ultimately have a very large file. The LINCS network will create access to the entire world's production of language information. The thesaurus with its ease of updating and adding index terms and relationships is an ideal index tool for very large files.

2) The scope of LINCS will be highly interdisciplinary, as the chart in Garvin's report suggests. The thesaurus can be structured to accommodate many points of view simultaneously.

3) LINCS will provide the necessary switching devices to interface with a variety of other information processing centers around the world, so that searches can be conducted for specific information indexed by different indexing languages. The thesaurus offers the requisite flexibility to switch from one index language to another.

4) Many of the contributors to Information in the Language Sciences [4] as well as Garvin [5], have alluded to the fact that language information, and especially "linguistics" is an emerging field with many schools and points of view, all of which must be accommodated in the information language of LINCS. LINCS obviously cannot be parochial. The thesaurus is uniquely able to structure index terminology in a "non-partisan" manner, to provide various hierarchical arrangements and cross references reflecting various views and taxonomies of the field. This capability is doubtless the single most important factor in our choice of the thesaurus as the indexing language for LINCS.

5) With the anticipated variety of users and uses for the information in the files of the LINCS network, a full range of generic to specific search capabilities must be provided by the index language. The index language must also provide, as far as is possible, for those unanticipated searches that will undoubtedly result from the interdisciplinary or "mission-oriented" use of LINCS. These requirements dictate the choice of: 1) an indexing language that can be post-coordinated at the time of the search, 2) an indexing language structured so that the term relationships are made evident to both indexer and searcher. The thesaurus will be designed to accomplish both of these tasks.

6) Although the ultimate configuration of the LINCS network cannot be precisely known at the outaet, an indexing language that is easy to use, both for indexer and searcher, must be provided, especially since it is possible that the input and searching processes will be performed at more than one location. An adequately documented thesaurus will be used in the same way by widely scattered indexers and searchers.

7) During the past decade, of all indexing tools, the structured controlled vocabulary known as the thesaurus has received the most sustained attention by information scientists. Its intellectual and physical structure has been the object of an enormous amount of effort culminating, perhaps, in the monumental Thesaurus of Engineering and Scientific Terms (TEST) of Project LEX [2].

TEST contains a very large and growing indexing vocabulary. It is designed to be used in an automated information retrieval system where all of the file searching and much of the thesaurus construction and maintenance is done by computer. Very sophisticated software has been developed for these purposes and is available to LINCS for experimentation.

The thesaurus offers the flexibility of structure and maintenance, the semantic controls and the cross referencing devices required for the LINCS indexing language by the considerations enumerated above.

4. Construction of Trial Thesaurus: Sources of Vocabulary

It is obvious that to be useful indexing must reflect the search needs of the user. Ideally, then, it might be proposed that an information storage and retrieval system should begin with the identification of its users, followed by the submission and collection of their own terminology for the indexing language. However, neither time nor money has permitted this purist approach in the past. Furthermore, we sincerely believe that

sufficient guidelines are now available as to potential users and potential scope of LINCS for us to begin preparation and pilot testing of a LINCS thesaurus. Conscientious evaluation studies of the system will ensure that the indexing tool of the operational LINCS will reflect the vocabulary and search objectives of its users. Hence, we are beginning to build the trial thesaurus using selections from the vocabulary sources described in Lewis [7] and Gifford [6], many of which have indeed been used to index the literature of linguistics.

5. Guidelines for Thesaurus Construction

A thesaurus is a controlled vocabulary to guide indexers and users. Its function is to bring the language of the author into coincidence with the language of the user who will be searching for information at some later time. How the thesaurus performs this function to a large extent determines the success or failure of an information retrieval system. It is therefore of the utmost importance that thesaurus makers know what they are doing, and that they lay down guidelines for processing terms, so that all decisions can be made consistently and in accordance with the purpose of the thesaurus.

Guidelines for thesaurus construction must deal with a wide range of problems from the most intensely philosophical to the purely mechanical. In drawing up the guidelines for constructing the LINCS Thesaurus we have used as a basis the USA Standard Basic Criteria for Indexes (USASI Standard) and the Guidelines for the Development of Information Retrieval Thesauri [1], prepared by the Committee on Scientific and Technical Information (COSATI). Our methodology will be that used by Project LEX to construct TEST. We are extremely indebted to such persons as Eugene Wall and others who have covered the same ground previously and left explicit instruction for thesaurus construction. All that remained for us to do was to adapt proven guidelines and methodology to the particular characteristics of linguistics.

6. Guidelines for LINCS Thesaurus Construction

Using the same forms developed for input of Terminology to TEST, we have prepared a sample of "core" terms displayed in thesaural relationships by the AUTO-LEX Thesaurus Construction and Maintenance Programs. An excerpt of this thesaurus is displayed in Chapter 4.

In order to use this form, the LINCS staff had to make decisions with specific reference to the language sciences on all the points listed in the COSATI Guidelines for thesaurus construction, as well as on some points not listed, but found from experience to be important. As was noted above, these decisions are a mixture of intellectual and clerical points. The list of points and the decisions made for the construction of the sample LINCS Thesaurus which will evolve into firm guidelines for LINCS are as follows:

1) Thesaurus Introduction

No introduction for the benefit of indexers and users has been written for this sample LINCS Thesaurus whose purpose is mainly to test the thesaurus programs and to display an array of linguistic terms in a thesaural structure.

2) Term Selection

The terms for the initail thesaural display were selected intuitively by linguists from available lists of indexing and vocabulary terms without specific data on their anticipated frequency in indexing or searching. They are all acceptable or authentic linguistic terms. Their relationships to other terms in the LINCS vocabulary is expected to change somewhat after more candidate terms are examined and after controlled indexing and searching experiments have been conducted.

3) Noun Form

The noun form of selected terms will be used in all instances where reasonable. For example, when we encounter the term parse, we shall enter it as the gerund, parsing.

4) Singular vs Plural

Although we have not done so in the sample thesaurus, we would probably be well advised to adhere to the rule of using plurals wherever possible. This rule would prevent the noun-verb ambiguity inherent in a term such as affix.

5) Term Ambiguity

We have tentatively attempted to clarify ambiguous terms by the use of parenthetical qualifying expressions, e.g., phonetics (acoustic), phonetics (articulatory), and phonetics (auditory). However, it is not yet resolved whether we shall ultimately clarify these kinds of ambiguity by qualifying notes in parentheses or by listing them as precoordinated terms, i.e., acoustic phonetics, articulatory phonetics, and auditory phonetics. In other instances we have freely included compound terms, such as anthropological linguistics. Specific guidelines for the use of one or the other, or both of these devices will be developed as more experience is gained.

6) Direct vs Inverted Entry

All terms except those in the language-name list are entered in LINCS directly without inversion, e.g., comparative linguistics, not linguistics, comparative. Whether or not uniform

guidelines should be established on this point has yet to be considered.

7) Synonyms

When two or more terms have appeared to be synonymous, we have selected one as the preferred term and entered the second as a USE reference, e.g., linguistic anthropology USE anthropological linguistics.

8) Punctuation

Except in the inversion of language names (Germanic, Western) punctuation has been avoided in the sample LINCS Thesaurus.

9) Abbreviated Word Forms

In the pilot vocabulary sample we have not encountered abbreviated word forms or acronyms, but we anticipate avoiding their use. For example, we shall use the term machine translation, not MT.

10) Alphabetization

The LINCS sample Thesaurus has been alphabetized according to the AUTO-LEX sorting program which is a (character-by-character) sort.

11) Cross References

The types of cross references as well as their notations used by TEST have been used in the LINCS Thesaurus. They are:

Type of cross reference	Notation
use	USE
used for	UF
broader term	BT
narrower term	NT
related term	RT

In the structured listing the main entry terms are displayed in alphabetical order down the left-hand column and the cross references are printed out beneath them indented to the right. The use of these cross references in the LINCS Thesaurus is as follows:

a. Use (USE) References

The USE reference leads the user of the thesaurus from a term that may be a valid term to the searcher or indexer to the term that is preferred by the thesaurus. It will be noted in this example that of the two terms historical linguistics and diachronic linguistics, both of which are "valid," in the general sense, the LINCS Thesaurus does not consider the latter a search term. Therefore the indexer and the searcher are given access to the thesaurus through both terms but are directed to use historical linguistics as their indexing or search term. USE is not optional, it is a directive. The term diachronic linguistics is not a LINCS term. It is anticipated that the USE reference will be very useful in the switching from one indexing vocabulary to another in the LINCS network.

It should be mentioned that the USE reference, while it may be used to indicate preference of one synonym over the other is not necessarily restricted to "pure" synonyms, but is used for those terms which are considered synonymous for indexing and retrieval purposes.

The U&E reference may also be incorporated in the language name part of the LINCS Thesaurus to lead the user, for example, from the more "hierarchically logical" Norse, Old to the operationally preferred term, Old Norse.

b. Used For (UF) References

The UF reference is the reciprocal of the USE reference and performs the same function of directing the user to the preferred LINCS term. For example, referring again to the terms discussed above, directly under the main entry historical linguistics is the entry UF diachronic linguistics, whereas directly under the main entry diachronic linguistics is the directive, USE historical linguistics.

The criterion for the selection of a USE or UF reference can in actual practice be almost arbitrary. They are both simply devices to control the index terminology of the thesaurus so as to

ensure that indexer and searcher entry vocabularies when they mean the same thing will be changed so that they can match in the search of the LINCS file.

c. Narrower Term (NT) and Broader Term (BT) References

These two references were developed to signify class inclusion relationships. Narrower terms are included in the meanings of broader terms, and broader terms include the meanings of narrower terms. It was in the attempt to develop explicit guidelines for the assignment of these references that the essential difference between linguistics and the "hard" sciences for which these references were originally developed became most apparent. The rule of thumb for thesaurus construction in the hard sciences is that a narrower term "is a" [member of the class] broader term. For example, steels are iron alloys would be designated by

steels
BT iron alloys

iron alloys
NT steels

However, linguistics is not a hard science. Its aspects partake of both humanities and the sciences, social and natural. Since it faces both ways, so to speak, this seemingly simple test for the NT-BT relationship is not feasible for the term arrangement of LINCS, except for some few terms denoting physical objects. On the other hand, the usual subjective way of arranging terms into a hierarchy which is usually expressed by "comes under" does not seem to be a proper criterion in the construction of a thesaurus. It would inevitably lead to the kinds of inconsistencies that make traditional library schemes so subject to criticism despite their attempts to adhere to principles of subdivision. Yet LINCS must develop a rule of thumb for consistent BT-NT relationships. The criterion which has been used to structure the terms of the sample thesaurus into BT-NT relationships is "if you were conducting a search for information indexed by the broader term, would you always want information indexed by the narrower term?" This criterion

is explicitly user-oriented and can only be validated by the users of LINCS. Later evaluation studies of LINCS will prove whether this guideline is viable. Of course, for the construction of the sample thesaurus, CAL was acting as user and indexer. The usefulness of this guideline can be illustrated by

allophone
  BT phoneme

and

phoneme
  NT allophone

which is to say that in a search for information on phonemes the user would always want information on allophones, but not necessarily vice-versa, because of an important policy of coordinate indexing, i.e., the indexer always assigns the most specific index term available. Thus while a user searching for information on phonemes would always want to see information on allophones, the user searching for specific information on allophones would not necessarily be interested in information about phonemes in general, or in any other aspect of phonemes. The search program provides for either kind of search.

The importance of using clear-cut, workable guidelines for indicating term relationships can not be over-emphasized. As the LINCS Thesaurus grows in size these guidelines will become more critical. If they are carefully developed and prove to be operationally feasible, they will ensure consistency of structure when terms are added to the LINCS Thesaurus which will in turn ensure consistency of search results.

d. Related Term (RT) References

The RT reference is used to refer from index terms to other index terms which are related, but not hierarchically, i.e., that are neither broader nor narrower. Since in the final analysis, every term in the file is related in some way, extreme caution should be exercised in assigning the RT reference. The fact that terms are indeed related in some unspecified way is not sufficient reason to indicate the RT relationship. The guideline for assigning this relationship

should be: "Would the user appreciate being reminded that the related term is available for searching?" We have used the RT reference sparingly in the LINCS sample thesaurus, and are not sure that it will be useful where we have used it, as, for example,

comparative linguistics
RT descriptive linguistics

The AUTO-LEX programs give the option of indicating the reciprocal of this relationship or suppressing it.

7. The Role of the LINCS Thesaurus in the LINCS Network

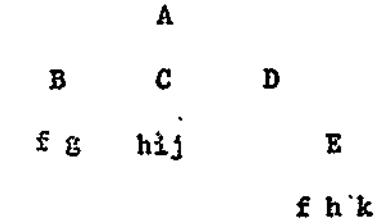
The LINCS Thesaurus promises to be particularly useful as a switching device in the LINCS network. If properly constructed it can be used to translate the various index languages used by the various centers comprising the network into the index language of LINCS. This capability becomes particularly important when one considers the bulk of material on the subject of the language sciences that is indexed in countries other than the United States, to which the LINCS network will give access. As an example of how the switching process might work between LINCS and a documentation center overseas using the Universal Decimal Classification to index language related information, we refer to the discussion by Robert Freeman on the subject. Freeman describes several potential solutions to the problem of gaining access to documents written in a foreign language to show the effectiveness of UDC to surmount language barriers:

> A third solution, which is attractive despite the greater effort which would be required to implement it, would be to permit indexing and searching to be done using a controlled natural-language vocabulary of local choice. A part of the system would then be a table of equivalences between the UDC and the natural language vocabulary. The result would be to take advantage of the hierarchical notation of the UDC without even requiring that the user be familiar with the UDC. In addition, since the UDC would be the internal form of indexing, users in any center could direct queries to the file, without regard to the original language in which the indexing was done. [3]

In Freeman's "third solution" the English "table of equivalences between the UDC and the natural-language vocabulary" could be incorporated into the LINCS Thesaurus in such a way that UDC numbers could be constructed by people, or possibly by computer, and searched by matching the request translated into UDC at the centers where UDC is used, thus avoiding the necessity of phrasing the query in a foreign language, or conversely, knowing the classification.

8. LINCS Microthesauri

As in any large information network where various member centers process specialized information, individual centers in the LINCS network will require more specific index terminology than will be useful for central LINCS. For these centers subsets of the LINCS Thesaurus can be extracted and used as a basis for more detailed microthesauri which will permit the specialized centers to index any desired specificity. These microthesauri will in turn be input to the internal LINCS Thesaurus so as to be available to all LINCS indexers should they need the specialized terms. We are tentatively planning to use the entire thesaurus including the microthesauri in the LINCS system to act as an internal devise to enable the indexing language to be controlled and standardized. The following tree is an illustration of how the microthesauri may be used as an internal control. Upper case letters represent terms in the "open" LINCS Thesaurus. Lower case letters represent terms in various microthesauri.



Note that terms f and h are placed in two separate hierarchical arrangments. With such a structure used internally, searches can be made for the specific terms f and h regardless of their hierarchical arrangment. At a more generic level, say B or E, questions can be negotiated to give the user the option of either hierarchy. This concept would give LINCS the flexibility it must have to various collections of language-related information indexed according to different points of view and taxonomies.

References

[1] Committee on Scientific and Technical Information (COSATI). Guidelines for the development of informational retrieval thesauri. Washington, D.C.: Government Printing Office, 1967.

[2] Department of Defense, Project Lex. Thesaurus of Engineering and Scientific Terms (TEST). Washington, D.C.: Government Printing Office.

[3] Freeman, Robert R. "Actual and potential role of the Universal Decimal Classification." In: Robert R. Freeman, Alfred Pietrzyk, and A. Hood Roberts, eds. Information in the language sciences. Proceedings of the conference held at Airlie House, Warrenton, Virginia, March 4-6, 1966, under the sponsorship of the Center for Applied Linguistics. Mathematical linguistics and automatic language processing 5. New York: American Elsevier, 1968, 149-163.

[4] Freeman, Robert R.; Pietrzyk, Alfred; Roberts, A. Hood, eds. Information in the language sciences. Proceedings of the conference held at Airlie House, Warrenton, Virginia, March 4-6, 1966, under the sponsorship of the Center for Applied Linguistics. Mathematical linguistics and automatic language processing 5. New York: American Elsevier, 1968, xi, 247 p.

[5] Garvin, Paul L. Specialty trends in the language sciences. LINCS #16-69, NSF GN-771. Washington, D.C.: Center for Applied Linguistics, 1969, ii, 29 p. [ERIC: ED 034 983].

[6] Gifford, Carolyn. A survey of indexing tools for the language sciences. CALLINCS-70P, NSF GN-771. Washington, D.C.: Center for Applied Linguistics, 1971.

[7] Lewis, Kathleen P., comp. Indexing tools and terminology sources in the language sciences: A bibliographical listing. LINCS #2-68, NSF GN-653. Washington, D.C.: Center for Applied Linguistics, 1968, 20 p. [ERIC: ED 021 245].

Chapter 2

VOCABULARY AND INDEXING FOR LINCS: SOME PRELIMINARY CONSIDERATIONS

By F.W. Lancaster

1. Requirements

The choice of indexing procedures and index language for LINCS will be dictated by: 1) the products and services to be provided, and 2) the organizational characteristics of LINCS itself.

LINCS will be a multipurpose system, generating a number of different products and services. Such products will probably include published indexes and abstracting journals, other current awareness devices including some form of SDI (on a group or individual basis), and retrospective search capabilities. It is important that the indexing and index language adopted should be capable of generating all of these products. That is, from a single input operation we must create an indexed data base from which all bibliographic services can be produced without further indexing modification. We do not want to index by one method for one service and a different method for another. Nor do we want the complication of having to produce complex algorithms to translate from one vocabulary to another (e.g., from a classification scheme to subject headings).

It is expected that LINCS will consist of a network (loosely structured) of information centers in the language sciences with both primary and secondary nodes. At the present time we expect that many of the operations of LINCS will be largely decentralized (as they are, for example, in the MEDLARS and ERIC networks). We expect to receive inputs (in the form of index records and/or abstracts) from several of these network components. The "LINCS Central" will be largely a network management center with responsibilities for policy, coordination, review, publication, quality control and network switching activities. Because network participation is likely to involve voluntary cooperative arrangements, indexing procedures are best kept relatively simple. We would like to avoid highly complex indexing methods or highly sophisticated indexing languages if the application of these would put an excessive burden on participating centers and thus tend to discourage full cooperation. Moreover, the LINCS network will incorporate information centers already in existence. Some of these components already produce document surrogates, of one type or another, for their own purposes. We would like to avoid duplication of effort by making use of these surrogates, intact or with minor modification, in the LINCS network as a whole. If necessary, we would want to convert from the vocabulary of an existing center (automatically or semi-automatically -- for example, by vocabulary conversion tables to allow mapping operations) to the vocabulary of LINCS, thereby allowing the center to

continue to index to meet its own specialized requirements but, at the same time, to be providing input compatible with LINCS requirements.

2. Vocabulary Alternatives

The following possible vocabulary approaches exist for consideration:

1) A carefully controlled, highly structured vocabulary in the form of a thesaurus, list of subject headings or classification scheme.

2) Free assignment of keywords or key phrases by indexers, as for example, in the technique of title expansion. Free use of keywords would perhaps be coupled with the use of some broad codes for subjects, countries, languages, etc.

3) Natural language searching and processing of abstracts, extracts or other document representation in machine-readable form.

4) Machine extraction of keywords or phrases.

5) Machine assignment of descriptors selected from a controlled vocabulary.

Before discussing the pros and cons of these various approaches, let us consider some general trends in vocabulary usage for information retrieval at the present time. It appears clearly that there is a general move toward simplicity in the exploitation of information retrieval systems. Such complexities as role indicators and similar syntactic devices are disappearing or are used very sparingly. The approach of natural-language searching, with comparatively little vocabulary control, is more popular now than previously. There are several reasons:

1) Experiments and operational experience have shown that natural-language systems can be made to work effectively.

2) Machine-readable corpora, by-products of photocomposition or various other keyboarding operations, are becoming widely available.

3) Natural-language searching is more attractive for on-line implementation than it was for batch-processing systems.

Various government agencies provide examples of the move toward simplification. Several years ago, certain major information systems utilized a highly sophisticated indexing, requiring skilled indexers and based upon a detailed classification scheme. Now these information systems use a relatively shallow indexing, based in some cases upon geographic codes, a broad and much abbreviated subject code (about 250 classes)

and uncontrolled keywords extracted from document titles or added to document titles. Indexing is now conducted by personnel with no more than a high school education, some of whom can index 100-125 documents per day. Indexing costs have thus been reduced dramatically. Further justification for systems based on some form of natural-language indexing and searching is provided by the following evidence:

1) In the comparison of index language devices conducted by the ASLIB Cranfield Project, it was clearly demonstrated that optimum retrieval results were achieved with natural-language and the simple device of term coordination. Only synonym control, and the confounding of word endings, improved on single-term natural-language searching. The more highly controlled "conceptual" index languages were out-performed by natural-language, single-term searching [1].

2) Salton, working with small experimental collections in several subject fields, has consistently produced acceptable results by fully automatic methods. In particular, the SMART-MEDLARS comparison suggests that automatic information systems, based on searching of natural-language abstracts, may now be able to perform as well as present-generation mechanized systems based on humanly-assigned index terms [3]. Salton's best results have usually been obtained with the less sophisticated of his search options.

3) Although no national information center has set up a retrieval system of this type, operating information systems based on natural-language do exist and have been shown to function effectively. Perhaps the most notable of these is the legal retrieval system established by Horty at the University of Pittsburgh [4]. These retrieval functions have now been taken over by the Aspen Systems Corporation.

Increased impetus to natural-language retrieval methods is given by the present availability of program packages for natural-language processing, including the IBM Document Processing System, which has been adopted by several large organizations.

With the foregoing background behind us, let us now consider the appropriateness of the various vocabulary alternatives for LINCS requirements.

Alternative 1 is perhaps the safest approach. Most large information services do make use of a structured, carefully controlled vocabulary. Such a vocabulary, in the form of a thesaurus or list of subject-headings, is capable of being used to produce the range of products planned for LINCS. MEDLARS, for example, uses a controlled vocabulary of this type and, from a single indexing operation, is able to produce printed indexes, demand searches and SDI service. It is relatively easy to achieve vocabulary compatibility when a controlled thesaurus is used.

Any specialized vocabularies existing in cooperating centers can become microthesauri within the framework of the general system thesaurus. This can be achieved by human mapping operations, leading to the production of machine-readable conversion tables. For example, the specialized vocabulary of the Parkinson's Disease Information Center has been mapped to the MEDLARS vocabulary of Medical Subject Headings. Some of the mapping may be done automatically if experiments with mapping algorithms, conducted by Wall, prove sucessful.

Once the vocabulary mapping has taken place, it is possible for the specialized center to index materials using its own vocabulary and indexing procedures but to have this indexing converted automatically to the vocabulary of the central system. Thus, one indexing operation serves both needs.

Another advantage of a fully controlled vocabulary is that, generally speaking, it improves search efficiency, reduces the burden on the searcher and may obviate the need for screening of system output before results are delivered to the user. The principal disadvantages are that the use of a controlled vocabulary (at least a large one) will usually lead to fairly expensive indexing (because of the look-up operations involved) and maintenance and updating of the vocabualry will also be a relatively expensive operation. Moreover, for efficiency, vocabulary control operations usually need to be centralized; decentralization can lead to many problems. A further possible disadvantage for LINCS is the fact that indexing using a large structured vocabulary is a relatively sophisticated operation requiring skilled indexers at the various participating centers. These indexers would need some training and also would be required to follow indexing rules and guidelines. These factors may reduce center tolerance to full participation in the LINCS network.

Alternative 2 is an attractive possibility. This would involve an indexing process whereby an indexer would assign some relatively broad subject codes, possibly some language or geographic codes, and several uncontrolled keywords. The keywords would probably be selected from the significant words occurring in titles plus additional significant words from the abstract or full text. These additional words may, in fact, be added by the indexer to the title to form an expanded title. Such indexing can be effected by a text-marking operation, as in the following example:

> (Mechanical) (Semantic Analysis) and the (Compatibility) of (English) (Adjectives) [(Protosyntrex III)]

in which each word or phrase enclosed within parentheses has been selected as a "keyword" and the expression enclosed in square brackets haa been added to the title to allow it to be picked up as an index term and also perhaps to clarify the title.

While the use of uncontrolled keywords alone can lead to much semantic ambiguity and noise, the joint use (as retrieval coordinates) of keywords

with broad subject and/or geographic codes produces a very powerful retrieval capability. The broad codes provide context for the keywords and reduce ambiguities. For example:

STRIKE associated with JORDAN

STRIKE associated with UNITED KINGDOM

If the former association occurs frequently it probably refers to a military context, the strike force. If the latter association occurs frequently it probably refers to a labor dispute.

The joint use of uncontrolled keywords and broad codes frequently allows a searcher to "zero in" on quite a small segment of a document file. For example, the strategy

LAMB (keyword) and AUSTRALIA and UNITED KINGDOM

geographic codes

will almost certainly retrieve documents relating to export of lamb from Australia to the United Kingdom and one cannot readily visualize much irrelevancy in this search.

This type of system, with an extremely large uncontrolled keyword vocabulary, is currently being used very successfully in retrospective search systems of major agencies whose document collections grow at the rate of about 250,000 documents per year. Such systems have shown to be feasible for SDI as well. It should also be suitable for LINCS published indexes, the broad subject categories being used for publication arrangement and the keywords for subject indexes.

For LINCS purposes this approach offers certain definite advantages. The approach, which is along the lines of procedures already used to produce indexes to such publications as The Finite String, should find ready acceptance at the various LINCS centers. Indexing is cheap and easy to accomplish and does not require an extensive investment in training programs and materials. The method is flexible enough to allow inputs in many different forms and from many different sources. It would be easy to integrate inputs from LINCS Central, LINCS Centers and many outside sources. There is no reason why relevant inputs from other information services (CFSTI, MEDLARS, for example) could not be incorporated into LINCS intact, using the indexing terms assigned by these centers as "keywords" in LINCS.

The problem of compatibility and convertibility between centers would be virtually eliminated if this approach were adopted. Further advantages are:

1) a highly specific, dynamic vocabulary reflecting current usage of terminology in the language sciences;

2) immediate implementation, without waiting for the completion of a thesaurus, and initiation of a training program.

Possible disadvantages are:

1) increased burden on searchers;

2) increased screening costs.

It should be noted that the use of an uncontrolled keyword vocabulary in indexing does not necessarily mean that no vocabulary control will be used in searching. Usually, some form of thesaurus or other logical grouping of terms will be needed to assist the searcher in construction of efficient search strategies.

Alternative 3, natural-language processing of abstracts, has also been proved (e.g., in SMART, in BROWSER developed by Williams of IBM) feasible for both retrospective search and SDI. However, some broad categorization scheme would still need to be employed as the basis for organization of abstracts in publications. The method is attractive for LINCS because a mechanism already exists for acquisition of abstracts, although not in machinable form. The production of abstracts may be more acceptable to centers than a formal indexing procedure. The language of the abstracts would yield a highly specific, dynamic vocabulary. Vocabulary maintenance costs need not be very high although some logical grouping of terms would be required to assist the searcher and improve search efficiency. Such program packages as the IBM Document Processing System exist to allow natural language searching of this type.

Implementation does require that abstracts be acquired for all items entering into the system and that these abstracts be put into machine readable form. However, it is likely that most LINCS publications would require the acquisition and keyboarding of abstracts in any case.

Alternative 4, machine extraction of keywords or phrases, has several of the advantages of Alternative 2. However, all programs for machine extraction (e.g., Klingbiel's) [2] are still experimental and no fully operating system exists to my knowledge. Moreover, many of the entry procedures for machine extraction (by statistical and/or syntactic criteria) have not been conspicously successful. Machine extraction involves the manipulation of at least an abstract in machine-readable form, so that we would not avoid this input cost.

If we go to the cost of capturing an abstract in machinable form, a term extraction procedure has little to commend it over free text searching of the complete abstract and requires much more complex and costly programming. This approach is definitely not recommended for LINCS at present.

Alternative 5, machine assignment of descriptors based on analysis of natural language text, is the most difficult to accomplish and has not been achieved very successfully in experiments thus far. It requires a machine-readable abstract, programming complications are increased, and the resulting retrieval system has less flexibility and specificity than one based on searching of natural-language text. This alternative is least attractive to LINCS at present.

On the basis of the above considerations it is obvious that at least three alternatives appear entirely feasible for LINCS implementation. All in all, however, considering the total LINCS requirement and in the light of our previous discussions on the subject, I am inclined to favor Alternative 2 as being probably least expensive and most readily implemented. The adoption of Alternative 2 at present does not preclude the possibility of switching to natural-language searching of abstracts at a later date (when the LINCS retrieval system is on-line and fully operational, say) if such a switch appears desirable. Indeed, it does not even preclude the possibility of switching at a later time to a fully controlled, structured vocabulary. In fact, the keyword vocabulary assembled in the uncontrolled indexing process will provide valuable raw material for continued thesaurus building. For this reason I favor continuance of work on the thesaurus. Some type of structured vocabulary will later be necessary as a searching aid in any event.

References

[1] Cleverdon, C.; Keen, M. Factors determining the performance of indexing systems. Vol. 2. Test results. Cranford, England: ASLIB Cranfield Project, 1966.

[2] Klingbiel, P.H. Machine-aided indexing. Alexandria, Va.: Defense Documentation Center, June 1969. [NTIS: AD 696 200]

[3] Salton, G. . A comparison between manual and automatic indexing methods. Ithaca, N.Y.: Cornell University, Department of Computer Science, March 1968.

[4] Springer, E.W.; Horty, J.T. Searching and collating the welfare laws of Pennsylvania by computer. Pittsburgh, Pa.: University of Pittsburgh, Health Law Center, September 1962. [NTIS: PB 164 437]

Chapter 3

A PRELIMINARY CLASSIFICATION FOR LANGUAGE SCIENCES INFORMATION: WORKING OUTLINE

By Fred Bauman

1. Introduction

There have been many classification schemes for linguistics; George Trager's 1945 scheme [4] is perhaps the most detailed of these, although others such as the linguistics sections of the Library of Congress Classification and the Universal Decimal System are much more actively in use. It is not the purpose of this outline, however, to discuss these classification systems; this work has already been performed by Carolyn Gifford in A survey of indexing tools in the language sciences [1], and adequate bibliographical references can be found in Kathleen P. Lewis' Indexing tools and terminology sources in the language sciences; a bibliographical listing [2]. This outline will, rather, first briefly discuss the pragmatic requirements for a classification system which could be used primarily as a framework for the thesaurus presently being prepared for the LINCS system, and then present a preliminary classification which attempts to meet some of these requirements.

Two important points about LINCS must first be made, because they influence the kind of classification system needed: 1) LINCS covers not just those fields which fall under a narrowly defined "linguistics" but rather the whole range of fields in which language is an important factor, i.e., the language sciences; 2) LINCS is an information network and as such must be primarily concerned with meeting the information needs of workers in the various fields of the language sciences. These two important factors influence both the scope and the structure of the classification system presented below.

Scope. Because the LINCS system attempts to cover the whole range of the language sciences, the classification system must include a wide range of fields. The present classification does this by an initial four part pragmatic division of the field into (1) Core Linguistics, which includes the traditional fields of linguistic endeavor; (2) Hybrid Linguistics, which includes those fields where linguistics interacts with another field of knowledge such as Sociology, Psychology or Mathematica; (3) Related Fields, which includes those non-language fields where developments may have important consequences for the Language Sciences; and (4) Languages.

Meeting User Needs. The second important point is that the classification system presented below is designed to meet the needs of the users of LINCS. The field of the Language Sciences has, accordingly, been defined not in terms of intellectually or theoretically established hierarchies but rather in terms of the literature in the language sciences in so far

as it reflects the work and the interests of researchers and scholars. Thus, the classification gives prominence to those fields which are prominent in the literature currently being produced.

Of great value in determining current fields of interest were the LINCS Reference Groups (see Chapter 4) and Priscilla Rose's Linguistic Bibliography Count [3]. The latter work was especially useful in deciding which fields in the classification required detailed hierarchical breakdowns. Thus, a field like onomastics, which in the 1966 Linguistic Bibliography was represented by only 38 entries, would not seem to require, for present purposes, the extensive breakdown provided by the Trager classification system, whereas fields like the Linguistic Bibliography's "Mathematical Linguistics," which is represented by 151 entries, would certainly seem to demand further breakdowns, such as provided in the preliminary classification outline presented below, where this area is covered by "Mathematical Linguistics" and "Language and Automation" and their subfields.

Response to user needs was also an important consideration in those instances where, because of the prominence of certain fields, they are given equal status with other fields to which they might actually seem subordinate. Thus "Teaching English as a Second or Foreign Language" might be thought of as a subgroup of "Foreign Language Education" but because of the importance of "Teaching English as a Second or Foreign Language" as reflected in the large number of publications in this area, it has been placed on the same level with "Foreign Language Education."

The chief features of the preliminary classification outline are, then, its broad scope, and its attempt to reflect the fields of the language sciences as represented in published literature. Since these are the requirements of the LINCS system, it is hoped that the present classification will be adequate to serve as a basis for work on the LINCS Thesaurus as well as for work on a more detailed classification for the language sciences.

2. Preliminary classification outline

[CORE LINGUISTICS]

THEORETICAL AND DESCRIPTIVE LINGUISTICS

Phonology
- Segmental Phonology
  - Phonetics
    - Acoustic Phonetics
    - Articulatory Phonetics
  - Phonemics
  - Distinctive Feature Analysis

Prosody [Suprasegmental Phonology]
Loudness, Stress, Amplitude
Timing (Length, Rhythm)
Pitch (Intonation, Tone)
Combinatory Phenomena (Emphasis, Juncture, Syllabification)

Grammar
Morphology
Syntax
Morphophonemics
Discourse (Analysis)

Lexicon
Lexicology and Lexicography
Etymology
Onomastics

Semantics
Structural Semantics
Semantic Theory

Orthography/Graphemics

CONTRASTIVE LINGUISTICS
Theories of Contrastive Linguistics
Error Analysis
Contrastive Analysis

COMPARATIVE AND HISTORICAL LINGUISTICS
Processes of Language Change
Language Reconstruction (Comparative Method)
Areal Linguistics

LANGUAGE CLASSIFICATION

LANGUAGE UNIVERSALS

LINGUISTIC THEORIES
Transformationalism
Stratificationalism
Tagmemics
Case Grammar
Prague School and Neo-Praguians
American Structuralism
Other

HISTORY OF LINGUISTICS

[HYBRID/HYPHENATED LINGUISTICS]

LANGUAGE AND BEHAVIOR

Theories of Verbal Behavior

Psycholinguistics
- Intellection
  - Cognition
  - Memory and Recall
- Child Language
  - Prelinguistic Vocalization
  - Development of Language in the Individual
- Psychoacoustics

Biolinguistics
- Neurolinguistics

Pathologies of Language Behavior
- Aphasia
- Non-aphasic speech pathology
- Non-aphasic dyslexia
- Psychopathology

Psycholinguistic Aspects of Bilingualism

LANGUAGE AND EDUCATION

Language Learning and Teaching (General)
- Theory of Language Learning/Teaching
- Physiology and Psychology of Language Learning
- Technology of Language Education
  - Audiovisual Techniques
  - Programmed Learning
    - Self-Instructional Techniques and Materials
    - Teaching Methods
  - Language Laboratories
  - Evaluation of Language-Learning Technologies
- Methodology (Other than "Technology of Language Education.")
- Teaching Materials (Other than "Technology of Language Education.")
- Language Testing
  - Achievement
  - Aptitude
  - Proficiency

Curriculum Studies
Teacher Education
Analysis and Teaching of Cross-Cultural Context

Foreign Language Education [See also "Language Learning and Teaching."]

Teaching English as a Second or Foreign Language [See also "Language Learning and Teaching."]

Native Language Teaching [See also "Language Learning and Teaching."
Language Arts
Social Dialects and Education
Standard Dialect for Speakers of Other Dialects

Bilingual Education

LANGUAGE AND SOCIETY

Sociology of Language [Fishman's Macrosociolinguistics]

National Language Situations
Language Planning
Language Policies
Language Standardization
Ethnic Minority Problems
Literacy
Bilingualism as a Group Phenomenon
Description
Theory
Languages in Contact
Diglossia
Bidialectism as a Group Phenomenon

Sociolinguistics [Fishman's Microsociolinguistics]

Social Dialect Description
Small Group Communication
Technical and Other Functional Styles
Bilingualism as an Individual Phenomenon
Description
Theory

LANGUAGE AND CULTURE [See also "Anthropology."]

- Linguistics and Anthropology
- Ethnolinguistics
- Ethnography of Communication

DIALECTOLOGY

- Linguistic Geography
  - Linguistic Atlases
- Dialect Descriptions

LINGUISTICS AND THE HUMANITIES

- Linguistics and Literature
  - Stylistics
  - Content Analysis
- Linguistics and Other Humanities

PHILOSOPHICAL LINGUISTICS

MATHEMATICAL LINGUISTICS

- Mathematical Models in Linguistics
- Quantitative Linguistics

LANGUAGE AND AUTOMATION

- Computational Linguistics
  - Automatic Language Processing
  - Computer Aids to Linguistic Analysis
  - Mechanical Translation
- Linguistics and Information Science
- Man-Machine Communication and Artificial Intelligence

TRANSLATION

SEMIOTICS

[RELATED FIELDS]

PHONETIC SCIENCES [See also "Phonetics."]

.PSYCHOLOGY

Cognitive Psychology [See also "Cognition."]

Developmental Psychology [See also "Development of Language in the Individual."]

Educational Psychology [See also "Language Learning and Teaching."]

Psychology of Perception [See also "Psychoacoustics."].

BIOLOGY [See also "Biolinguistics."]

Speech Physiology
Hearing Physiology

MEDICINE AND THERAPY

EDUCATION

SOCIOLOGY

Socioeconomic Studies

ANTHROPOLOGY [See also "Language and Culture."]

Cognitive Anthropology
Social Anthropology

POLITICAL SCIENCE

Ethnic Minority Problems

GEOGRAPHY

Demography

MATHEMATICS

COMPUTER SCIENCE

INFORMATION PROCESSING AND DOCUMENTATION

INFORMATION AND COMMUNICATION THEORY

PHILOSOPHY

HUMANITIES

- Literature
- Music

LANGUAGES*

INDO-HITTITE MACRO-PHYLUM

Anatolian Family

Indo-European Phylum

- Albanian Isolate
- Armenian Family
- Baltic Family
- Celtic Family
- Germanic Family
- Hellenic Family
- Illyrian Family
- Indic Family
- Iranian Family
- Italo-Romance Family
- Slavic Family
- Tocharian

URALIC-ALTAIC MACRO-PHYLUM

Uralic Phylum

- Finno-Ugric Family
- Samoyedic Family

*The outline classification for Languages was prepared by Charles Zisa.

Altaic Phylum

- Korean Isolate
- Mongolian Family
- Tungusic Family
- Turkic Family

AFRO-ASIATIC MACRO-PHYLUM

- Berber Family
- Chadic Family
- Cushitic Family
- Hamitic (Egypto-Coptic) Family
- Semitic Family

AUSTRALIAN MACRO-PHYLUM

SINO-TIBETAN MACRO-PHYLUM

- Kam-Thai Family
- Sinnitic Family
- Tibeto-Burman Phylum

AUSTRONESIAN MACRO-PHYLUM

AFRICAN LANGUAGES

- Niger-Congo Phylum
  - Adamawa-Eastern
  - Central (Bantu)
  - Gur
  - Kordofanian
  - Kwa
  - Western Atlantic
- Nilo-Hamitic Family
- Nilo-Saharan Phylum
  - Chari-Nile
  - Sudanic
- Khoisan (Bushman-Hottentot) Phylum

AMERICAN INDIAN LANGUAGES

- Algonquian Macro-Phylum
- Andean-Equatorian Macro-Phylum
- Azteco-Tanoan Phylum
- Chibchan Macro-Phylum
- Ge-Pano-Carib Macro-Phylum
  - Hokan Phylum
  - Na-Dene Phylym
  - Oto-Manguean Phylum
- Siouan Macro-Phylum
- Ungrouped Amerindian Languages and Groups

CAUCASIAN LANGUAGES

- North Caucasian Phylum
- South Caucasian Family

PAPUAN LANGUAGES

SOUTHEAST ASIAN LANGUAGES

- Andamanese Languages
- Jakunic Family
- Sakaic Family
- Salweenic Family
- Semangic Family
- Vietnamic Family

BASQUE FAMILY

DRAVIDIAN FAMILY

ESKIMO-CHUKCHEE PHYLUM

MUNDA FAMILY

NIPPONIC (JAPANESE-OKINAWAN) FAMILY

PALEO-SIBERIAN PHYLUM (AINU-GILYAK, KET, YUKAGHIR)

PIDGIN AND CREOLE LANGUAGES

UNGROUPED LANGUAGES

References

[1] Gifford, Carolyn. A survey of indexing tools for the language sciences. CALLINCS-70-6, NSF GN-771. Washington, D.C.: Center for Applied Linguistics, 1971.

[2] Lewis, Kathleen P., comp. Indexing tools and terminology sources in the languages sciences: A bibliographical listing. LINCS #2-68, NSF GN-653. Washington, D.C.: Center for Applied Linguistics, 1968, 20 p. [ERIC: ED 021 245]

[3] Rose, Priscilla. Linguistic bibliography count. LINCS #10-70, NSF GN-771. Washington, D.C.: Center for Applied Linguistics, 1971.

[4] Trager, G.L. "A bibliographical classification system for linguistics and languages." Studies in Linguistics, 1945, 3:54-108.

Chapter 4

VOCABULARY CONTROL FOR THE LINCS REFERENCE MANAGEMENT SYSTEM (RMS)

By Alfred Pietrzyk

This outline summarizes the initial indexing approaches and authority file management techniques which, at this time, are considered to be optimal for use in the proposed Reference Management System (RMS), the automated central clearinghouse and secondary processing facility of LINCS. Figure 1 shows the general configuration of the envisaged RMS. Most of the modules (1-6) will in some way be effected by coordinated vocabulary control techniques. The emphasis of this outline is on Module 6 for authority file management. If these plans are actually implemented, several modifications will no doubt turn out to be desirable.

1. Indexing

Human indexing at the input processing stage (Module 1) will be dynamic: standard terms from the RMS authority files (thesaurus descriptors and language names, broad subject category terms, and auxiliary terms) will be used in conjunction with identifiers, i.e., current natural language terms and context-preserving phrases based directly on the source information and/or its surrogates. Reference units (document surrogates) will be indexed to an average of 8-10 terms.

2. Authority File Management

2.1 Baseline

With a view toward effective vocabulary control for reference materials in the language sciences, the LINCS program has completed important preliminaries, with the following overall findings and results:

- The indexing philosophy of LINCS must be dynamic, i.e. both controlled and uncontrolled open-ended vocabulary must be used at the human indexing stage in a carefully combined approach in order to ensure
  - high recall in search operations by using controlled generic thesaurus terms and controlled broad subject category terms;
  - high precision in search operations by using both controlled specific thesaurus terms and uncontrolled specific terms and phrases extracted from natural language text;
  - compatibility with structured indexing tools of cooperating information processing and service organizations;

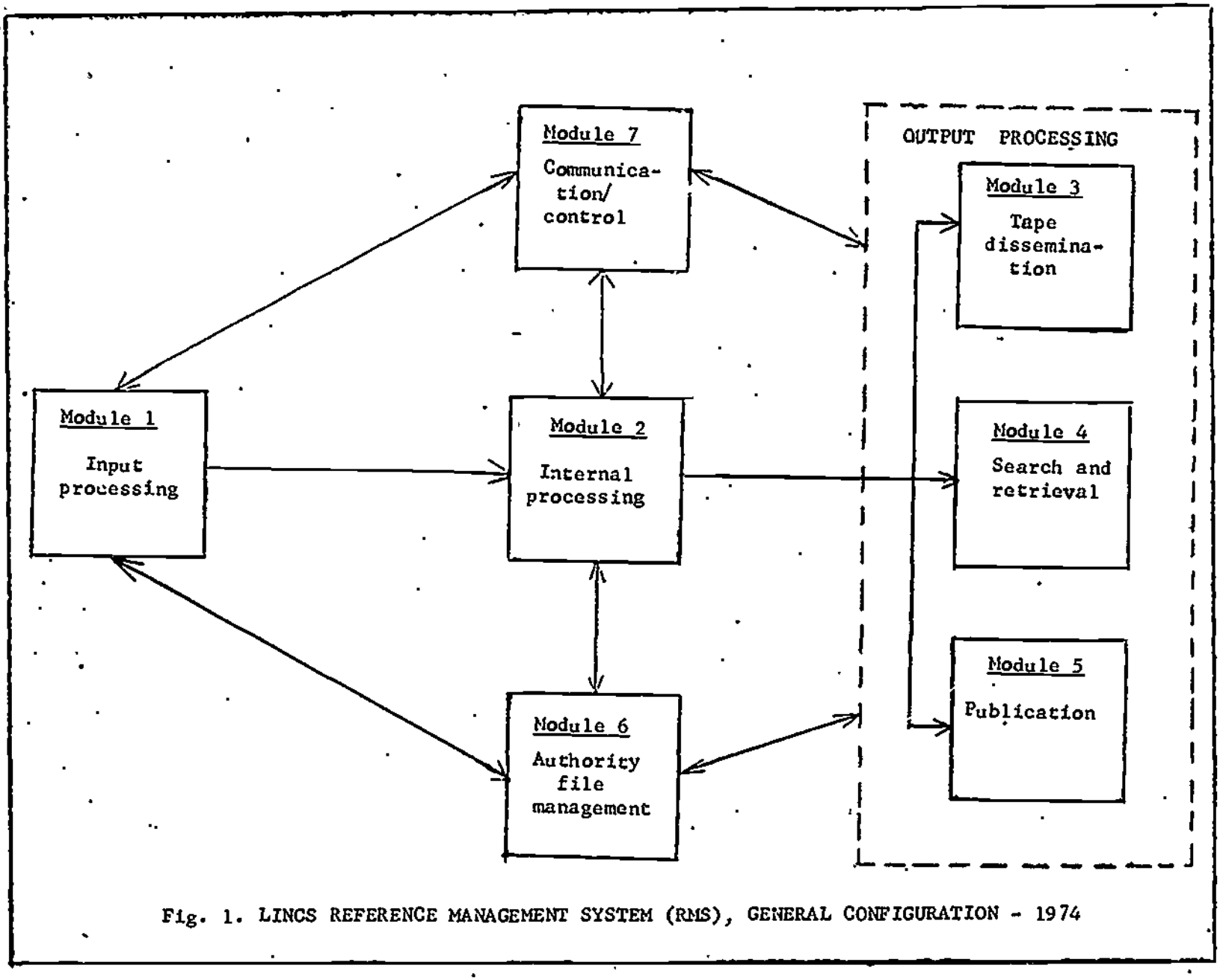


Fig. 1. LINCS REFERENCE MANAGEMENT SYSTEM (RMS), GENERAL CONFIGURATION - 1974

- currency of the indexing vocabulary;

- comprehensive coverage;

- preservation (in indexing phrases) of syntactic contexts with high information content.

- The following preliminary drafts of indexing tools and source materials have been completed or acquired:

  - an experimental sample thesaurus for LINCS, prepared by Joy Varley of the LINCS staff; the thesaurus contains some 450 unique technical terms (descriptors) in the language sciences, with considerable specificity in one subfield (phonology), structured in accordance with COSATI guidelines, including items under USE, USED FOR (UF), BROADER TERM (BT), NARROWER TERM (NT), RELATED TERM (RT), and SCOPE NOTE (SC), with hierarchical display of narrower terms to a depth of five levels (see Figure 2);

  - a preliminary classification outline (see Chapter 3);

  - a comprehensive coded list of some 5,000 unique language and dialect names (17,000 entries including synonyms) prepared by the CAL for NSF's National Register of Scientific and Technical Personnel (the codes cover generic sets);

  - a detailed classification of American Indian languages (954 unique items, 3,730 entries including synonyms);

  - a listing of some 190 broad subject category terms under 46 reference group headings (see Table 1);

  - 18 controlled auxiliary terms describing document type and status (e.g. "dictionary," "revision");

  - a comprehensive collection of existing thesauri, microthesauri, technical dictionaries, indexes, and classifications relevant to the language sciences, usable as source materials for thesaurus construction (not suitable for direct use in the proposed RMS)

- A limited capability for automated thesaurus display (see Figure 2) has been assembled on an experimental basis.

```
WHISPERED VOWELS
  BT   VOWELS
       SEGMENTAL PHONEMES
       PHONEMES
       PHONOLOGY
  RT   WHISPER


WORD CLASSES
  UF   PARTS OF SPEECH
  BT   DESCRIPTIVE (STRUCTURAL) LINGUISTICS
    *NT  ADJECTIVES
         NOUNS
         VERBS


WRITING
  SC   THE MECHANICS OF WRITING
  BT   LITERACY
    *NT  ORTHOGRAPHY
           SPELLING
             SPELLING REFORM
         SOUND-SPELLING CORRESPONDENCES


WRITING SYSTEMS
  USE ORTHOGRAPHY
```

Fig. 2. LINCS THESAURUS EXCERPT (UNEXPANDED PRELIMINARY DRAFT)

A series of LINCS reports deals with preliminaries to thesaurus construction and maintenance, indexing options, and classification principles (see References, Part Two, final LINCS project report).* Significant practical experience was gained in the machine-aided production of permuted subject indexes for the experimental reference serial Language and Automation.

For purposes of the proposed RMS, the following requirements remain unfulfilled:

- All authority files must be improved, modified, and integrated to accommodate precisely all human and automated processing requirements in RMS Modules 1-5, including requirements for compatible interfaces with decentralized collaborators.

- The LINCS thesaurus must be refined and expanded to achieve comprehensive coverage of technical terms (descriptors) as well as language and dialect names needed in RMS processing (the current draft version does not include language names).

- The authority files for broad subject category terms and auxiliary terms must be improved and expanded for comprehensive coverage.

- Human and automated procedures for authority file construction and maintenance must be fully specified.

- The existing limited automated aids for thesaurus processing must be re-designed for the increased, more complex requirements listed above, in order to ensure accurate and prompt maintenance of all authority files needed in the RMS. The current automated capability is uneconomical; the proprietary program now in use cannot be modified to include the required input/edit/update functions and additional display formats required minimally for efficient authority file management.

- The initial design of integrated authority file management approaches must be open-ended for future automation refinements (see Figure 3).

2.2 Objective for 1974

Module 6 will be an operationally ready subsystem for computer-supported authority file management in the language sciences, with

*Center for Applied Linguistics. An information system program for the language sciences: Final project report, NSF Grant GN-771. CALLINCS-71-4. Washington, D.C.: Center for Applied Linguistics, 1971.

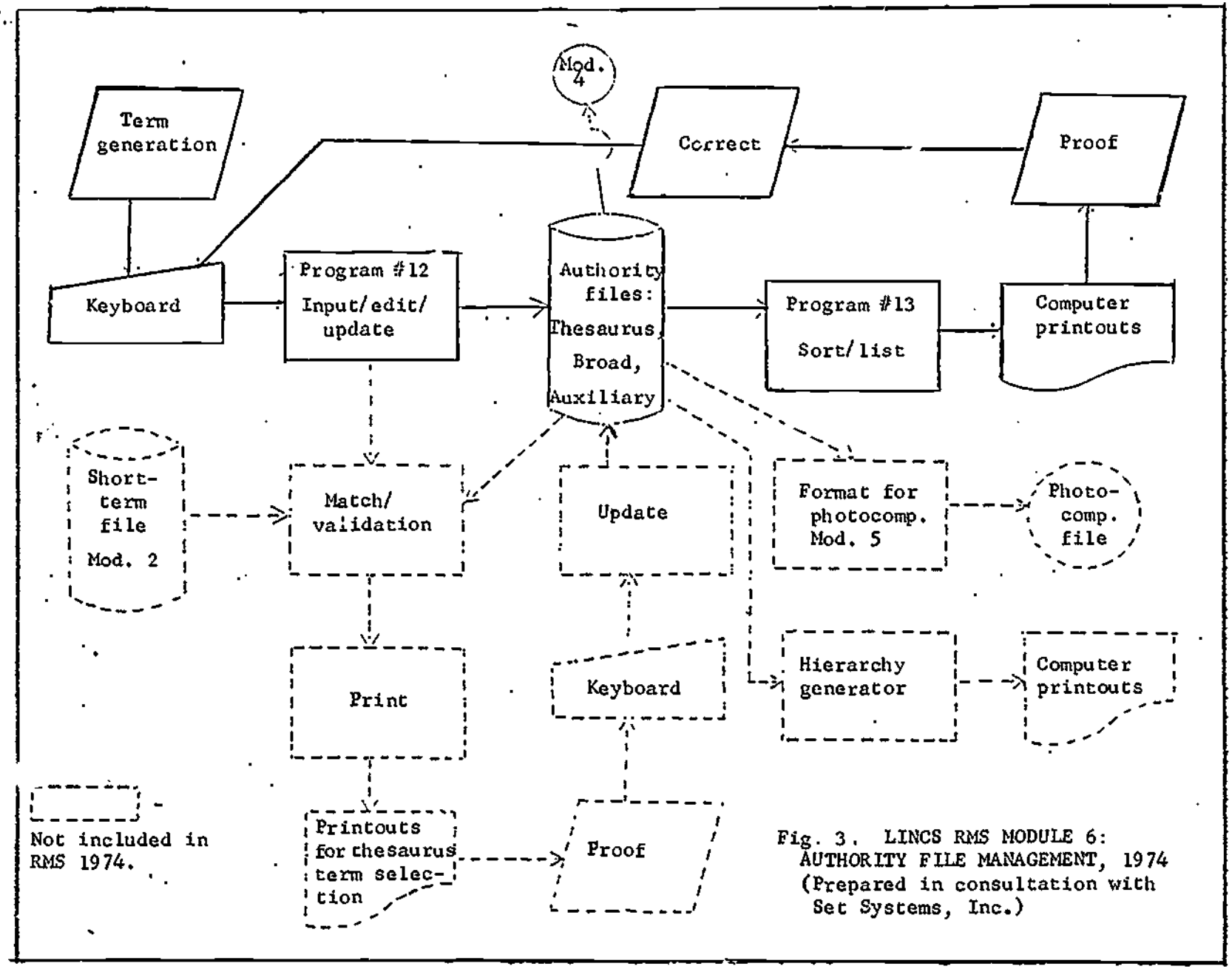


Fig. 3. LINCS RMS MODULE 6: AUTHORITY FILE MANAGEMENT, 1974 (Prepared in consultation with Set Systems, Inc.)

specific application to vocabulary control needs of the RMS and its input and output processing interfaces (Figure 3). Its main functions will be:

- to provide comprehensive desk-top tools (periodically updated computer printouts of authority files) for vocabulary control - including thesaurus control - in centralized and (standardized) decentralized human indexing at the input processing stage (Module 1);

- to provide machine-readable terms needed for automated validation of broad subject category terms and auxiliary terms in reference file maintenance (Module 2);

- to provide vocabulary control - including thesaurus control - in the formulation of search strategies (Module 4);

- to provide vocabulary control - including thesaurus control - in the extracting, index, and sortiug operations of the publication subsystem (Module 5);

- to utilize human operations for the continuous maintenance of all RMS authority files (term generation and structuring, keyboarding, editing, updating, proofing, correction, and preparation of desk-top tools and tapes for use in other RMS modules);

- to provide a minimal, economical capability for continous computer-supported processing and maintenance of the thesaurus, broad subject category terms, and auxiliary terms (including input/edit (validate)/update functions, master file storage, and sort/list/print in thesaurus and other formats required by the RMS).

Additional functions will be added after 1973 (see Figure 3).

Following their initial construction in the RMS project, all authority files will be maintained in regular update cycles. Intellectual efforts will concentrate on term extraction from current sources and term structuring in accordance with COSATI guidelines modified for RMS purposes. The following improved, fully expanded machine-readable authority files will be available by 1974:

- a comprehensive, structurally refined language sciences thesaurus based on COSATI guidelines (cf. Figure 2) containing about 5,000 unique descriptors (technical terms) and about 5,000 unique language and dialect names;

- a comprehensive file of about 400 improved broad subject category terms used in packaging of outputs and cooperative exchanges of inputs (these broad terms will also be included in the thesaurus);

- about 30 improved auxiliary terms (initial sample only by 1974) used as document type and status descriptors.

The results will include full specifications of human and automated procedures for authority file construction and maintenance. The module assembly will include an economical computer facility (IBM 360/30).

The authority file capability will consist of the following main functional flow components (Figure 3):

- intellectual processing of new or revised authority file terms; term collection from current sources, visual matching against existing files, structuring and formatting for automated input processing;

- off-line keyboarding of term inputs on magnetic tape recording typewriter (including off-line machine-aided proofing and correction);

- computer input and partial machine validation of new terms, also error listing and maintenance of processing statistics (RMS program #12);

- updating of machine-readable master files for thesaurus, broad terms, and auxiliary terms (program #12);

- sorting of master file subsets and display (printout) including thesaurus format, alphabetical listing, and permuted term format (program #13);

- proofing, correction and re-entry of corrected terms via keyboarding and input/edit/update components;

- preparation of partial or comprehensive machine-readable files for use in RMS file maintenance (Module 2) and output processing (Modules 3-5).

Table 1. REFERENCE GROUPS IN THE LANGUAGE SCIENCES*

The 46 user-oriented reference groups listed below have been established on the basis of operational criteria including the productivity of published research in given areas. Pragmatic criteria prevail over intellectual and taxomonic principles. Together, the reference groups cover the entire spectrum of the language sciences. The subject categories given for certain reference groups are illustrative rather than exhaustive. The subject categories are listed approximately in accordance with their relative importance in a reference group. Certain categories of primary importance in one reference group re-occur as secondary categories in other reference groups. The general linguistics group cuts across the entire set of reference groups. However, services in this category involve, in part, a non-overlapping subset of the total audience. The number of potential LINCS users in 1976 estimated for each reference group includes only those users with a primary interest in the group involved, i.e., all figures listed are non-overlapping. Specific services focused on various reference groups will, of course, be offered to wider audiences. Likewise, the number of message units (articles, books, etc.) expected in 1976 has been estimated in each case only for material of focal interest. Given services will, however, include selections from other reference groups. The reference group concept is dynamic; it will be continously refined and modified in the light of changing user requirements, advice from the community, and newly evolving research and publication patterns.

Reference Group	Total no. of users, 1976	Total no. of message units (articles, books, etc.) 1976
1 GENERAL LINGUISTICS	18,150	2,870
History of linguistics		
Theoretical linguistics		
Descriptive linguistics		
Historical linguistics		
Other language sciences		
All language groupings		

* Prepared in collaboration with Joy Varley and other members of the LINCS staff, as well as consultants specializing in various subfields of the language sceinces.

Reference Group	Total no. of users, 1976	Total no. of message units (articles, books, etc.) 1976
2 PHONETIC SCIENCES	9,320	1,680
Acoustic phonetics		
Physiological phonetics		
Perceptual phonetics (speech perception)		
Descriptive phonetics		
Historical phonetics		
Statistical phonetics		
Phonology/phonemics		
Automatic speech analysis and synthesis		
Phonetics and communication sciences		
Psychoacoustics		
Phoniatrics		
Logopedics		
3 THEORETICAL AND DESCRIPTIVE LINGUISTICS	12,870	2,100
Foundations of linguistics		
"Schools" of linguistics		
Theory of phonology		
Theory of writing		
Theory of grammar		
Semantic theory		
Language universals		
Formal and mathematical linguistics		
Linguistic methodology		
Descriptive linguistics (principles)		
Historical linguistics (principles)		
Linguistic phylogeny		
Linguistic ontogeny		
Typology of languages		
Linguistics and logic		
Linguistics and philosophy		
Other language sciences		
History of linguistics		

Reference Group	Total no. of users, 1976	Total no. of message units (articles, books, etc.) 1976
4. LEXICOLOGY AND LEXICOGRAPHY	2,365	390
Lexical theory and applications Monolingual dictionaries Bilingual dictionaries Bidialectal dictionaries Multilingual dictionaries Etymological dictionaries Bilingualism Specialized terminologies General thesauri Information retrieval thesauri Lexical planning Etymology Automatic dictionary lookup Automatic dictionary publishing Theoretical and descriptive linguistics		
5. HISTORICAL LINGUISTICS AND CLASSICAL LANGUAGES	4,830	640
Diachronic linguistics (theoretical and descriptive) Comparative method Glottochronology Lexicostatistics Proto-language reconstruction Classical languages		
6 LINGUISTIC GEOGRAPHY	3,220	960
Dialectology Linguistic atlases Dialect descriptions Censuses Onomastics Bilingualism		

Reference Group	Total no. of users, 1976	Total no. of message units (articles, books, etc.) 1976
7 LANGUAGE AND CULTURE	11,590	2,390
Linguistics and anthropology Ethnolinguistics Cognitive anthropology Ethnographic semantics Ethnography of communication Sociolinguistics Speech communities Area studies Culture history Language and mission work Literacy		
8 SOCIAL DIALECTS AND EDUCATION	11,020	1,410
Microsociolinguistics Social dialect description Bidialectalism Psycholinguistics Small group communication Ethnic minority dialects Standard dialects for speakers of other dialects Technical and other functional styles Social anthropology Social psychology Socioeconomic studies Sociology		
9 LANGUAGE PROBLEMS AND LANGUAGE PLANNING	4,050	790
Macrosociolinguistics National language situations Language planning Language codification (standardization) Linguistic innovation and borrowing Orthography Orthoepy Language policies		

	Reference Group	Total no. of users, 1976	Total no. of message units (articles, books, etc.) 1976
	Literacy Language maintenance and shift Ethnic minority problems Bilingualism Multilingualism Specialized terminologies Languages of wider communication Second language learning Artificial languages Pidgins and creoles		
10	BILINGUALISM	11,690	1,240
	Bilingualism theory Bilingualism description Languages in contact Contrastive linguistics Diglossia Multilingualism Bidialectalism Linguistic borrowing Language and culture Psycholinguistics Language problems and language planning		
11	CONTRASTIVE LINGUISTICS	8,000	1,040
	Theory of contrastive linguistics Contrastive analyses Error analyses Bilingualism		
12	FOREIGN AND SECOND LANGUAGE EDUCATION	29,540	3,530
	Language teaching methodology Physiology and psychology of language learning Technology of language education		

Reference Group		Total no. of users, 1976	Total no. of message units (articles, books, etc.) 1976
	Language ability testing		
	Teacher education		
	Teaching materials		
	Curriculum studies		
	Program evaluation		
	Language aptitude testing		
	Analysis and teaching of the cross-cultural language context		
	Psycholinguiatics		
13	TECHNOLOGY OF LANGUAGE EDUCATION	10,215	1,410
	Audiovisual techniques		
	Programmed learning		
	Self-instructional techniques and materials		
	Teaching machines		
	Language laboratories		
	Tape collections		
	Evaluation of language-learning technologies		
	Psycholinguistics		
	Language and culture		
	Language and automation		
14	LANGUAGE AND BEHAVIOR	13,885	2,780
	Psycholinguistics		
	Verbal behavior		
	Linguistics and cognitive psychology		
	Neurolinguistics		
	Psychoacoustics		
	Language and the child		
	Biolinguistics		
	Pathology of language		
	Psychology of perception		
	Psychology of learning		
	Developmental psychology		
	Psychometrics		
	Educational psychology		
	Special education		

Reference Group	Total no. of users, 1976	Total no. of message units (articles, books, etc.) 1976
15 LINGUISTICS AND MEDICINE	33,355	4,470
Speech physiology Speech pathology Hearing physiology Hearing pathology Aphasia Dyslexia Neurolinguistics Language and mental health Biolinguistics Psychiatry Psychopathology Phoniatrics and logopedics Otolaryngology Audiology and audiometrics Human communication disorders Communication of the blind Medical terminology Language education of the handicapped		
16 LINGUISTICS AND THE HUMANITIES	4,730	700
Language and literature Linguistics and philology Linguistics and poetry Stylistics Rhetoric Stylostatistics Content analysis Classical and mediaeval studies Linguistics and music Linguistics and other humanities Language and culture Mass communication		
17 LANGUAGE AND AUTOMATION	12,960	1,660
Computational linguistics (automatic language processing) Quantitative linguistics		

	Reference Group	Total no. of users, 1976	Total no. of message units (articles, books, etc.) 1976
	Mechanical translation Machine-aided language learning Linguistics and computer science Theoretical and descriptive linguistics Automation in the humanities and social sciences Artificial intelligence Man-machine communication		
18	SEMIOTICS	5,210	1,300
	Theory of signs Paralinguistics Proxemics Kinesics Human communication Animal communication (zoosemiotics) Ethology Anthropology		
19	TRANSLATION	15,490	1,600
	Human translation theory Human translation applications Theory of machine translation Machine-aided translation Lexicology and lexicography Dictionaries Specialized terminologies Sociolinguistics		
20	ONOMASTICS	1,160	250
	Anthroponymy Toponymy Lexicology and lexicography		

Reference Group		Total no. of users, 1976	Total no. of message units (articles, books, etc.) 1976
21	FRENCH	21,050	2,770
22	IBERIAN LANGUAGES	21,400	2,740
23	ITALIAN	3,515	550
24	ENGLISH LINGUISTICS	25,720	3,120
25	ENGLISH AS A NATIVE LANGUAGE	28,100	3,940
26	ENGLISH FOR SPEAKERS OF OTHER LANGUAGES	12,640	1,750
27	GERMAN	9,470	1,410
28	SCANDINAVIAN	1,475	560
29	SLAVIC AND BALTIC	2,005	510
30	LANGUAGES OF THE SOVIET UNION	2,520	490
31	RUSSIAN	6,660	1,080
32	URALIC	960	150
33	ALTAIC	1,115	150
34	SOUTH ASIAN	1,875	240
35	SOUTHEAST ASIAN	3,290	500
36	CHINESE	6,560	1,240
37	JAPANESE	5,075	950
38	AFRO-ASIATIC	4,730	810
39	LANGUAGES OF SUB-SAHARAN AFRICA	6,700	950
40	MALAYO-POLYNESIAN	1,250	250

Reference Group		Total no. of users, 1976	Total no. of message units (articles, books, etc.) 1976
41	PACIFIC LANGUAGES	1,245	230
42	AUSTRALIAN LANGUAGES	1,010	150
43	NORTH AMERICAN INDIAN; ESKIMO AND ALEUT	1,915	300
44	SOUTH AMERICAN INDIAN	2,780	650
45	PIDGINS AND CREOLES	1,245	130
46	ARTIFICIAL AND AUXILIARY LANGUAGES	8,505	1,040
		406,460	59,870